Support for High Priority Traffic Using Preemption

Hongchuan Xiao and Cory Beard
University of Missouri-Kansas City

Abstract--To support high priority traffic, low priority traffic must either be restricted or preempted when high priority traffic arrives. This work provides a detailed study and optimization of resource control schemes using preemption. Mechanisms for computing blocking and preemption probabilities are provided, then objective functions are proposed to optimize relative costs and profit. These optimization objectives are then used to investigate network conditions where preemption policies would be best, or where other mechanisms would be better given the cost of preempting connections that are in progress.

Index terms-- preemption, priority, emergency

I. INTRODUCTION

Future broadband networks are being designed to integrate all types of multimedia traffic. More importantly, however, they are being designed to integrate and support the activities of all types of users. As networks become more widely deployed and have higher performance capabilities, new types of user applications will emerge and important types of uses will become more prevalent. As people become familiar with these applications, they will come to rely on them and expect networks to make them available when needed.

A user type of particular interest is the National Security/Emergency Preparedness (NS/EP) user. NS/EP users currently use the public Internet for outreach, information sharing, and electronic mail. However, use of the public Internet for mission-critical activities is currently modest [1], [2]. Instead of using the public Internet for critical functions, the NS/EP community depends more on dedicated TCP/IP networks. The reliability and security of the public Internet is considered inadequate for mission-critical functions, even though several useful applications have been identified that could benefit from use of the public Internet [3].

The purpose of this work is to propose preemption mechanisms as an alternative to extensive research that has been conducted on traditional connection admission control (CAC) policies (see, for example, [4], [5] and references therein). Previous work can be classified as developing two categories of policies.

1. First come-first served - Users are allocated resources simply if resources are available.

2. Reservation - Some resources are reserved for high priority users (or resources are restricted for low priority users) so that high priority users have a better opportunity to obtain resources when they need them (i.e., have lower blocking probabilities).

Each of these types of policies has its own drawbacks. The first category lacks a mechanism that can deal with different classes of users, which is not efficient in circumstances where contention for resources is strong. The second category wastes resources when high priority users are not offering much traffic to the network, because the capacity reserved for them goes unused.

To avoid these disadvantages, preemption can be used. The concept is to allow all users into the network whenever resources are available, then interrupt and discontinue the flows of lower priority users if high priority users need resources but there is no room to accommodate them.

Intuitively, preemption is a good idea. Little research has been conducted to study preemption, however, especially from a performance and optimization standpoint compared to other alternatives. One reason why the study of preemption has not been conducted is the preemption seems to be a harsh approach. People would rather block at the beginning of a session than be interrupted during a session. Because preemption is seen to be so onerous, two of the major initiatives in the United States to support high priority users, the Government Emergency Telecommunications System (GETS) [6] and the International Emergency Preparedness Scheme (IEPS) [7], [8], explicitly choose not to allow preemption. In contrast, however, military networks use multi-level precedence and preemption which does support interruption of lower priority users [9].

While objections to preemption are justified, preemption still should be considered a viable alternative. The main issue of preemption is dissatisfaction of low priority users. Given appropriate incentives, however, low priority users would certainly be amenable to preemption if they were provided, for example, rebates or lower subscription rates. As a matter of fact, in times of emergency it would likely take little to placate a low priority user if they were preempted; they would just need to know that they were preempted so an emergency worker could use the network.

Once one agrees that it is possible to eliminate (or at least offset) irritation to low priority users, preemption is a problem worthy of significant investigation. Even if preemption were acceptable, it must be studied to determine if it would really provide any benefit. The following sections are organized as follows. First preemp-
tion is defined and analyzed through an analysis of multi-dimensional Markov chains. Then optimization problems are formulated to find the best preemption policies to maximize profit (revenue minus rebates to preempted users) or minimize relative costs of blocking and preemption. And finally, scenarios are investigated to compare preemption against traditional complete sharing (CS) and upper limit (UL) policies.

II. DEFINITIONS AND BACKGROUND

The aim of this work is to determine whether preemption can be used in an overloaded network to improve the performance of the network in terms of both technical and economic criteria. Furthermore, if the answer is positive, a follow-on goal is to determine the magnitude of the improvement and whether the improvement justifies the application of preemption.

In other words, the goals of this paper are to answer the following questions.

1. Does preemption have advantages over other policies for managing and allocating capacity to network flows? Preemption presents a large inconvenience to lower priority connections that are preempted, which must be offset using economic mechanisms through rebates or some other approach. Here we compare preemption to CS and UL policies which do not use preemption.

2. How practical is implementation of preemption in a real network environment? How strong of an economic benefit does preemption provide?

It should be clarified that preemption here is a mechanism to be implemented for an entire flow or connection; it does not relate to packet-level scheduling mechanisms. Preemption can be classified according to the following three definitions.

- **Hard preemption** - Flows are interrupted and discontinued.
- **Soft preemption** - Flows are given a reduced quality of service level to free some of the flow’s resources for a higher priority flow. Flows do not end.
- **Delayed preemption** - Various time delays are used in conjunction with hard preemption. For example, flows can be interrupted and return later or can be notified of impending preemption but given a short time period before it occurs.

This work focuses on hard preemption as the first stage of a larger project to consider all three types of preemption. The goal is to find blocking and preemption probabilities in a preemptive system, optimize blocking and preemption to minimize cost or maximize profit, and see if a preemptive system has distinct advantages over non-preemptive approaches.

Preemption has been studied to some degree, but primarily in the context of signaling requirements or methods for choosing which connections to preempt. In [10], an algorithm was developed for adding preemption functionality to the connection establishment process in ATM networks. In our own work [11], extensions were made to provide specifications for Q.2931 signaling messages and the message processing procedures that need to occur. In general, the implementation of preemption from a signaling standpoint requires new messages or message contents, new handling and hand-shaking procedures, and use of timers if forms of delayed preemption are used. If hard preemption is used, technically no signaling modifications are necessary, since users could just be disconnected, but signaling to indicate such actions would certainly be beneficial.

When one connection is to preempt one or more other connections, procedures have been developed for choosing which connections to preempt [12], [13], [14]. Network operators may choose to seek to preempt the fewest number of connections, fewest amount of bandwidth, or minimize the cumulative priority levels. The presence of future knowledge can greatly help in the selection of the connections to preempt, but efficient online algorithms are possible. Here no consideration is made about which connections are to be preempted; it is just assumed that connections are chosen randomly.

III. PREEMPTION AND BLOCKING PROBABILITIES

Before computing preemption and blocking probabilities for a system using hard preemption, it is first important to precisely define the probability of preemption. This can be defined in two ways.

1. \( P_{\text{call}} \): The probability that a flow from class \( i \) is interrupted, out of *all* the flows of class \( i \) that arrive into a system = \( \Pr\{\text{preemption} \mid \text{an arrival}\} \).

2. \( P_i \): The probability that a flow is interrupted, out of *flows* of class \( i \) that are admitted into a system (i.e., not blocked) = \( \Pr\{\text{preemption} \mid \text{admission}\} \).

The definition of \( P_i \) seems to be more intuitive and similar to what one would naturally assume when considering the concept of preemption. This work will use \( P_i \) for most calculations.

To find \( P_i \), first of all the following assumptions are made. Flows for class \( i \) arrive according to a Markov process of rate \( \lambda_i \) and have a Markov service process of rate \( \mu_i \). All classes of traffic share a single link of capacity \( C \) basic bandwidth units, a flow from class \( i \) requires \( b_i \) units of equivalent bandwidth on the link, and each class has an infinite population. The network is assumed to be in an overloaded state where traffic levels create significant levels of congestion and blocking of resource requests. It is assumed that a higher priority flow will preempt one or more lower priority flows ran-
randomly, without respect to time in progress or time until finishing.

Consider the case of 2 classes, where class 2 has preemptive priority over class 1, Figure 1 shows the two dimensional Markov chain that results. This illustrates a system with $C=6$, $b_1=1$, and $b_2=3$. Arrival and departure transitions are at rates consistent with non-preemptive 2-dimensional Markov chains, with arrival rates of $\lambda_i$ and service rates out of a state with $n_i$ connections in progress of $n_i\mu_i$. In addition, diagonal transitions indicate preemptions. For example, the line from state $(n_1=5, n_2=0)$ to state $(3,1)$, indicates a preemption where an arriving class 2 flow preempts two class 1 flows.

Figure 1 is an example of what we call "full preemption". In such a case, all preemptions that could occur are allowed. If some transitions were not allowed, then this would be called "partial preemption". An example of partial preemption would be to say that a class 2 flow is allowed to preempt one or two class 1 flows, but not three. In such a case, Figure 1 would not have transitions from $(6,0)$ to $(3,1)$ or from $(3,1)$ to $(0,2)$. In most of the optimization cases presented in Section V, one finds an optimum solution with either full preemption or no preemption, but there are cases where a partial preemption policy is optimal.

To find preemption probabilities, it is first necessary to find state probabilities for the Markov chain. This is found by solving for equilibrium probabilities for a continuous time Markov chain using the following equations,

$$ \mathbf{pQ} = 0, \sum_{i=1}^{N} p_i = 1, $$

where $\mathbf{p}$ is the vector of state probabilities, $\mathbf{Q}$ is the matrix of transition rates, and $N$ is the number of states.

A state can be described by a state vector $\mathbf{n}=(n_1,n_2)$, the state probability for that state is $\pi(\mathbf{n})$, and a set $\Omega$ is defined as the set of states from which preemptions can occur. For example, for Figure 1, $\Omega = \{(1,1), (2,1), (3,1), (4,0), (5,0), (6,0)\}$. The set of states from which blocking can occur for class $i$ is $\Phi_i$. The function $f(\mathbf{n})$ is defined as $f(\mathbf{n})=1$ if a preemption is allowed from a state and $f(\mathbf{n})=0$ if it is not. And finally, the number of class 1 flows that must be preempted (and lost) from state $\mathbf{n}$ to allow for one more class 2 flows is $l(\mathbf{n})$.

When a class 2 flow arrives, the expected number of class 1 flows it will preempt, $\alpha$, is as follows.

$$ \alpha = \text{expected number of class } 1 \text{ flows that will be preempted upon an arrival of a class } 2 \text{ flow} $$

$$ \alpha = \sum_{\mathbf{n} \in \Omega} f(\mathbf{n}) \pi(\mathbf{n}) $$

Then the average rate at which class 1 flows will be preempted is $\alpha \lambda_2$ and $P_{1,all}$ is the ratio of the average rate at which class 1 flows are preempted divided by the average rate at which class 1 flows arrive,

$$ P_{1,all} = \frac{\alpha \lambda_2}{\lambda_1}. $$

Since $P_1 = \Pr\{\text{preemption | admission}\}$,

$$ P_1 = \frac{\Pr\{\text{preemption} \cap \text{admission}\}}{\Pr\{\text{admission}\}} = \frac{P_{1,all}}{1-B_1} $$

$$ = \frac{\lambda_2 \sum_{\mathbf{n} \in \Omega} f(\mathbf{n}) \pi(\mathbf{n})}{\lambda_1 (1-B_1)} $$

where $B_1$ is the blocking probability for class 1,

$$ B_1 = \sum_{\mathbf{n} \in \Phi_1} \pi(\mathbf{n}). $$

A newly arriving class 1 flow has three possible fates.

1. Out of all the class 1 flows that arrive, some are blocked with probability $B_i$.
2. Others are preempted before they finish, with probability $P_{1,all}$.
3. The rest finish normally, with probability $1-B_1-P_{1,all}$.

$P_i$ is the conditional probability of preemption of those that are admitted. This can be viewed as the ratio of item #2 to items #2 and #3 as follows.

$$ P_1 = \frac{P_{1,all}}{1-B_1-P_{1,all}+P_{1,all}} = \frac{P_{1,all}}{1-B_1} $$

![Figure 1 - 2-Dimensional Markov Chain for a Full Preemption Policy](image-url)
IV. OPTIMIZATION FORMULATIONS

The fundamental goal of this work is to find whether preemption can improve the performance of networks. The most important indices to be used to make this determination are relative cost and profit. These indices can be used to find an optimal preemptive policy. Either no preemption, full preemption, or partial preemption will be best. Then the optimum result can be compared to the optimum result that could be produced for non-preemptive policies.

In the relative cost model, relative weights assigned to preemption events and blocking events. These weights are used to compute a weighted sum of blocking and preemption. Then this weighted sum is minimized. In the profit model, each flow is assumed to provide revenue to the network provider. To compensate low priority users for the nuisance of being preempted, however, network providers also provide a rebate to customers if they are preempted. The relative cost model uses less parameters, so is less dependent on the specifics of a particular implementation. A profit model is more detailed and seeks to more closely replicate an actual network. Each are described in the following subsections.

First of all, the following parameters are defined.

\[ C_{bi} = \text{Cost of blocking a flow from class } i \text{ for one unit of bandwidth.} \]
\[ C_{pi} = \text{Cost of preempting a flow from class } i \text{ for one unit of bandwidth.} \]
\[ T_{pi} = \text{Per-connection rebate provided for each preempted connection.} \]
\[ R_{ti} = \text{Time-based revenue generated from a flow from class } i \text{ for one unit of bandwidth.} \]
\[ R_{ci} = \text{Connection-based revenue generated from a flow from class } i \text{ for one unit of bandwidth.} \]

Connection-based revenue is generated when a connection is admitted. Time-based revenue is generated as the time of a call increases.

A. Minimum Relative Cost

The relative cost formulation is similar to the weighted sum of blocking formulation used in [4]. For a two-class system with class 2 having preemptive priority over class 1, relative cost is as follows.

\[ \text{Cost} = \lambda_1 b_1 (C_{b1} B_1 + C_{pi} P_{1,all}) + \lambda_2 b_2 (C_{b2} B_2) \]  

(5)

To find minimum cost, one then would consider various preemption policies that would provide different balances between \( B_1, B_2 \), and \( P_{1,all} \). Cost parameters could be absolutely determined according to an economic analysis of the services being offered by a service provider.

This model also, however, allows for a relative cost assignment, by setting \( C_{b1}, C_{b2} \), and \( C_{pi} \) with values relative to each other. For example, one could set \( C_{b1} = 1 \), then say the cost of blocking a class 2 flow is twice that \( (C_{b2} = 2) \), and the cost of admitting a flow but preemtting it before it is finished is \( C_{pi} = 3C_{b1} = 3 \).

The real determination of the usefulness of preemption lies in the relationship between \( C_{pi} \) and \( C_{b2} \). If \( C_{pi} \) is very large, it would be best to do as little preemption as possible (even no preemption). If \( C_{pi} \) is small (even to the limiting case of \( C_{pi} = 0 \)), then it would be best to do as much preemption as possible, since the negative impact of preemption would be small.

Consider the following example illustrated in Figure 2. If we have a system with \( C = 30, b_1 = 1 \), and \( b_2 = 10 \), then a class 2 flow may need to preempt up to 10 class 1 flows. Given \( \lambda_1 = 1.5, \lambda_2 = 0.15 \) \((b_1 \lambda_1 = b_2 \lambda_2)\), \( C_{b1} = 1 \), and \( C_{b2} = 2 \), then Figure 2 shows the plots of three curves for various values of \( C_{pi} \). For \( C_{pi} = 1 \), the minimum relative cost (indicated by an asterisk) is where a preemption policy would allow up to 10 class 1 flows to be preempted (i.e., full preemption). For \( C_{pi} = 3 \), the optimal policy would allow up to 5 class 1 flows to be preempted (partial preemption), and for \( C_{pi} = 15 \), none should be preempted (no preemption).

This example shows that partial preemption can be an important method for providing minimal relative cost. For the range \( C_{pi} \leq 1.9 \), full preemption is optimal; for \( C_{pi} \geq 12.8 \), no preemption is optimal. For a very large range, however, \( 1.9 < C_{pi} < 12.8 \), partial preemption is best. When two classes have very different bandwidth requirements, we expect that partial preemption will frequently yield an optimum result, because it provides a compromise between extremes of no preemption and full preemption.

For \( R \) classes, relative cost would be
\[
\text{Cost} = \sum_{i=1}^{R} \lambda_i b_i \left( C_{bi} B_i + C_{pi} P_{i,all} \right),
\]
and \( P_{i,all} \) would be zero for the highest priority classes.

**B. Maximum Profit**

The minimum relative cost formulation can be used in cases where only the three cost parameters are known or when relative magnitudes of those three parameters are to be considered. When a more detailed economic model is available, a maximum profit formulation can be used.

Profit is defined as revenue minus cost. Revenue can be generated from flows both when a flow is first admitted (using a per-connection charge \( R_{ci} \)) and as a flow progresses using a per unit time charge, \( R_{ti} \). Rebates are provided for those flows which are preempted on a per-connection basis (a per unit time rebate does not seem to have much usefulness).

In [15], equation (1), a formula was developed for profit in a preemptive system. It made the same assumption that is used here that a per-connection rebate was given to preempted calls, but it also assumed that no revenue was collected from preempted calls. Here we propose a more general scenario where preempted calls provide revenue to a service provider, which is then counteracted with a rebate. Depending on how long a connection had already been in progress before the preemption, a customer may net an amount greater than or less then zero. If a connection had been in progress for a long time, a customer may still provide some revenue to the service provider, but less than would have been provided. When a customer is preempted, the service provider loses the time-based revenue they would have received, plus they lose the amount of money provided in the rebate.

The equation for profit is then created as follows. First consider two classes where once again class 2 connections preempt class 1 connections. The average duration of a class 1 connection is \( \frac{1}{\mu_1} \). The average duration of those connections which are preempted, designated by \( d_p \), is biased by the process of selecting which flows to preempt. Note that \( d_p \) is meant to designate the time requirement of the flow (if the job had been allowed to finish), not the elapsed time at the time a job is preempted. The average duration of those connections which are not preempted is denoted as \( d_{np} \). These two durations are related through

\[
d_{np} (1-P_1) + d_p P_1 = \frac{1}{\mu_1}.
\]

The elapsed time can be found from the assumption of a Markov service time process. When a flow is interrupted it is expected to have its expected service time still remaining because the Markov process is memoryless. Therefore, the average amount of time a class 1 call has been in progress before it is preempted (i.e., the part that generates revenue) is

\[
d_p - \frac{1}{\mu_1}.
\]

The average profit for class 1 is the sum of profit from non-preempted class 1 connections and preempted class 1 connections as follows.

Class 1 Profit =

\[
\lambda_1 b_1 (1-B_1) \left( R_{c1} + R_{t1} d_{np} \right) + \lambda_1 b_1 (1-B_1) P_1 \left( R_{c1} + \frac{R_{t1}}{\mu_1} d_p - \frac{1}{\mu_1} T_{p1} \right). \]

Total profit across both classes is then

Total Profit =

\[
\lambda_1 b_1 (1-B_1) \left( R_{c1} + \frac{R_{t1}}{\mu_1} \right) (1-P_1) - P_1 T_{p1} + \lambda_2 b_2 (1-B_2) \left( R_{c2} + \frac{R_{t2}}{\mu_2} \right). \]

Total profit is discussed in more detail when it is used in Section V to compare optimal preemption policies to other types of non-preemptive policies.

**C. Equivalency of Minimum Relative Cost and Maximum Profit**

The two optimization formulas given in (11) and (6) can produce the same optimal preemption policy if parameters are chosen as follows. Profit can be viewed as the following

Total profit =

\[
(\text{Maximum profit with zero preemption and zero blocking})
\]

minus (Cost of lost revenue from blocking)

minus (Cost of lost revenue from preemption)

minus (Cost of rebates for preempted flows)

Maximum profit is

\[
\lambda_1 b_1 \left( R_{c1} + \frac{R_{t1}}{\mu_1} \right) + \lambda_2 b_2 \left( R_{c2} + \frac{R_{t2}}{\mu_2} \right). \]
The cost of lost revenue from blocking is analogous to the parameter $C_{bi}$ for class $i$, 
\[ C_{bi} = R_{ci} + \frac{R_{ti}}{\mu_i}. \]  
(13)

The parameter $C_{bi}$ then corresponds to the average amount of revenue that is expected from each class $i$ call. The cost of preemption, $C_{p1}$, is a combination of the cost of lost revenue and the cost of rebates,
\[ C_{p1} = \frac{R_{t1}}{\mu_1} + T_{p1}. \]  
(14)

It should be noted, however, that the cost parameters in the relative cost model could also include intangible factors that are of importance to a service provider from a management or public relations standpoint. For example, a service provider may choose to use a higher value of $C_{p1}$ than is given from (14) to indicate that the effect of preemption on end-users is even greater than can be captured from lost revenue and the cost of rebates.

To extend the example in Section IV.A and Figure 2 with specific values, the average connection duration for both classes is $1/\mu_1 = 1/\mu_2 = 10$ minutes. To have the given values of $C_{bi} = 1$ and $C_{b2} = 2$, $R_{ci}$ can be chosen to be $R_{c1} = $0.50 and $R_{c2} = $0.05 per minute. $R_{c2}$ can be chosen to be $R_{c2} = $1.00 and $R_{t1} = $0.10 per minute.

Full preemption is best if $C_{p1} \leq 1.9$, which would correspond to a rebate $T_{p1} \leq $1.40. For $C_{p1} \geq 12.8$, no preemption is optimal, which would correspond to $T_{p1} \geq $12.30.

D. Special Case of Equal Bandwidths

Before proceeding with comparisons of preemptive policies with other types of policies in Section V, it is helpful to consider the special case for situations were classes 1 and 2 have equal bandwidth requirements, $b_1 = b_2 = b$. It is reasonable to assume that such cases would arise, since no significant reason exists to say that a high priority connection would use any more bandwidth than a low priority connection.

Through observation it was found that in such cases the blocking for class 1, $B_1$, is constant regardless of the type of preemption policy that is used, whether it be full, partial, or no preemption. From this it also can be shown that
\[ P_1 = \frac{\lambda_2 (B_1 - B_2)}{\lambda_1 (1 - B_1)}. \]  
(15)

which means that $P_1$ is a linear function of $B_2$ (since $B_1$ is constant). Profit can be then written as a sum of constants plus a term that depends on $P_1$,

Total Profit = (constant terms) + $P_1(\lambda_2 b)(1 - B_1) \left( \frac{R_{t2}}{\mu_2} + \frac{R_{c2}}{\mu_2} - \frac{R_{t1}}{\mu_1} - T_{p1} \right).$  
(16)

The optimal policy is then dependent on the following term, since all other terms are constants.
\[ \frac{R_{t2}}{\mu_2} + \frac{R_{c2}}{\mu_2} - \frac{R_{t1}}{\mu_1} - T_{p1} \]  
(17)

If (17) equals zero, then profit is constant, regardless of the value of $P_1$, so all types of preemption policies yield the same result. This means that constant profit results in a system of equal bandwidths where the average per-connection revenue for class 2 connections (first two terms of (17)) is equal to the average rate-based revenue for class 1 that is lost for each preempted connection plus the rebate amount given to each of those preempted connections (as was seen in (14)). Note that the per-connection revenue rate for class 1, $R_{c1}$, is not a factor.

If (17) is greater than zero, which would occur when the possible revenue from class 2 is large, the maximum value of $P_1$ that can be obtained will produce the maximum profit. This would occur with full preemption. And finally, if (17) is less than zero, then the value $P_1 = 0$ would produce the maximum profit, which would occur in the case of no preemption.

V. COMPARISONS

The optimal preemption policy for a given set of network conditions can be found by minimizing relative cost or maximizing profit. The overall goal of this work is to see how optimum preemption policies compare with other types of policies. In this section policies will be compared based on maximizing profit.

The first type of policy to consider is a complete sharing (CS) policy. In a CS policy, all connections are admitted simply if resources are available at the time a connection is requested. A CS policy does not consider the importance of a connection when resources are allocated. A CS policy is a limiting case of a preemptive policy with no preemption.

The other type of policy to consider is an upper limit (UL) policy. In a UL policy, lower priority classes are limited in the amount of resources they are allowed to use in the network [4]. An upper limit policy is one example of a set of similar policies, like trunk reservation [5], virtual partitioning [16], etc., which limit access to resources by lower priority classes. The main concern about these policies is that they reduce the utilization and revenue generation potential of a network. They reduce the amount of load that can be accepted from low priority users for the sake of providing reduced blocking probabilities to high priority users.
Generally speaking, the final decision about whether a preemptive, CS, or UL policy yields an optimum result is case-dependent. Some general trends do exist, however. For overloaded networks, we have found that UL policies are better than preemption policies for moderate values of $R_{t2}$. When $R_{t2}$ is large (meaning class 2 connections are very important or generate a large amount of revenue), however, preemption policies are best. This can be illustrated as follows. First it is important to develop a scenario where the effects of UL, CS, and preemption policies are clearly seen without undue influence from particular choices of parameters. For this reason we consider an overloaded system with equal bandwidths ($b_1=b_2=1$), $C=50$, equal service rates ($\mu_1=\mu_2=0.5$), zero per-connection charges ($R_{c1}=R_{c2}=0$), and equal arrival rates ($\lambda_1=\lambda_2=25$). The goal is to consider the effect of the revenue rate for class 2, $R_{t2}$, when the larger this rate the more important it is to favor class 2 using preemption or a UL policy. The class 1 revenue and cost parameters are $R_{t1}=1$ and $T_{p1}=2$.

Using (17), a full preemption policy is better than a CS policy if

$$ R_{t2} \geq \mu_2 \left( \frac{R_{t1}}{\mu_1} + T_{p1} - R_{c2} \right) $$

Comparison with the optimal UL policy is slightly more complicated. Figure 3 shows how the optimal UL policy is different for various values of $R_{t2}$. An asterisk shows the point where maximum profit is achieved. For the curve with $R_{t2}=0.9$, maximum profit is achieved when $L_1$ equals the capacity of the system, which is really a CS policy with no restrictions on class 1.

The curve for $R_{t2}=1.5$ shows the effect of an additional constraint. For UL policies with high values of $R_{t2}$, the optimal policy would be to have $L_1=0$, which would not allow any traffic for class 1 to be admitted (i.e., $B_1=1$). This is not acceptable, however, so Figure 4 illustrates the effect on blocking for various values of $L_1$. Figure 3 uses an optimization constraint that $B_1$ must be less than 0.9, so for $R_{t2}=1.5$, the optimum is affected by that constraint.

Figure 5 shows a comparison of all three policies. For the range $R_{t2} \leq 1.0$, CS and UL policies provide maximum revenue. In the range $1.0 \leq R_{t2} \leq 5.5$, an upper limit policy is best, and for $R_{t2} \geq 5.5$, a preemption policy is best. This illustrates that in this case a preemption policy is better than a CS policy as soon as more revenue is expected from a class 2 connection that is expected to be lost due to preemption. To be better than a UL policy, however, a preemption policy must generate much more revenue (over two times as much) than is lost.

VI. CONCLUSIONS

This work produced mathematical constructs for com-
puting state probabilities in a loss system that also allowed high priority connections to conduct hard preemption of in-progress low priority connections. These state probabilities were used to compute blocking and preemption probabilities, as well as find optimal policies with regard to minimizing relative cost or maximizing profit (revenue minus rebates given to preempted customers). Optimal preemptive policies were then compared to optimal UL policies and CS policies. Preemptive policies and were found to be useful in cases where revenue from high priority connections was expected to be high. In addition, from a relative cost standpoint from Section IV.A, partial preemption was shown to be very useful in cases of unequal bandwidths.

Several opportunities exist to extend this work and provide a full picture of the relevance and usefulness of preemption. We believe algorithms could be developed to compute blocking and preemption probabilities without inversion of the Q matrix. Observations have been made that B1 is constant in the case of equal bandwidths, and for arbitrary bandwidths that B2 is the same for a UL policy with L1=0 and the corresponding full preemption policy.

Only hard preemption has been studied in this work, and the merits of soft preemption and delayed preemption are yet to be determined. This work should also be extended to multiple links and an arbitrary number of service classes. Work will continue in an effort to give a complete picture of the merits of preemption and its ability to support the needs of high priority users who have urgent needs for communications resources. High priority uses of broadband networks are certain to increase, and preemption should be understood and applied where it can be most helpful.

REFERENCES